Review and Outlook

Machine Learning 2019
mlvu.github.io

the plan

part 1:

Review

Exam strategies

part 2:

Outlook: the future and impact of machine learning

week 1
Introduction
Classification, Regresssion, Clustering
Linear Models 1
Hyperplanes, Random Search, Gradient descent

week 2
Method 1
Comparing methods
Method 2
preprocessing
missing values, outliers, dim. reduction

week 3
Probability 1
Gaussians, (Naive) Bayes
Entropy, Logistic regression
Linear Models 2
Neural nets, SVMs

week 4
Deep Learning 1
SGD
Backpropagation
CNNs
Probability 2
Expectation Maximization
!pick a topic!

week 5
Deep Learning 2
Generative models:
GANs, VAEs
Tree Models & ensembles
Decision trees
Boosting, Bagging

week 6
Sequences
Markov models, Word2Vec, RNNs
Matrix models
Recommender systems
PCA
Graph Models

week 7
Reinforcement Learning
Q learning, Policy gradients, AlphaGo/Zero/Star
Review
Open problems
The future of machine learning

week 8
exam
25 March

project deadline
29 March

introduction

Abstract (part of) your problem to a **standard task.**

Classification, Regression, Clustering, Density estimation, Generative Modeling, Online learning, Reinforcement Learning, Structured Output Learning

Choose your **instances** and their **features**.

For supervised learning, choose a target.

Choose your **model class**.

Linear models, Decision Trees, kNN,

Search for a good model.

Usually, a model comes with its own search method. Sometimes multiple options are available.

linear models 1

pick a random point **m** in the model space

loop:

$\mathbf{m} \leftarrow \mathbf{m} - \eta \nabla \text{loss}(\mathbf{m})$

we usually set η somewhere between 0.0001 and 0.1

methodology 1

training	validation	test

During model and hyperparam. selection:

- train on: **training**
- test on: **validation**

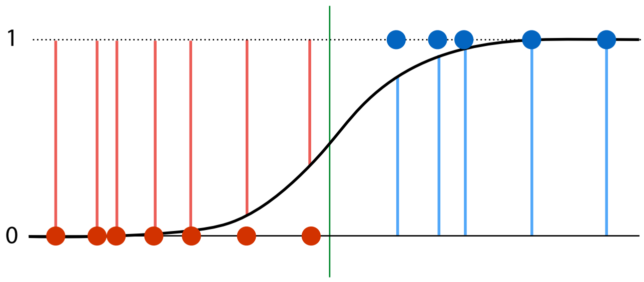
Final run:

- train on: **training** **validation**
- test on: **test**

Think about the REAL-WORLD use case.

probabilistic methods 1

$$c(x) = \sigma(\mathbf{w} \cdot \mathbf{x} + b)$$



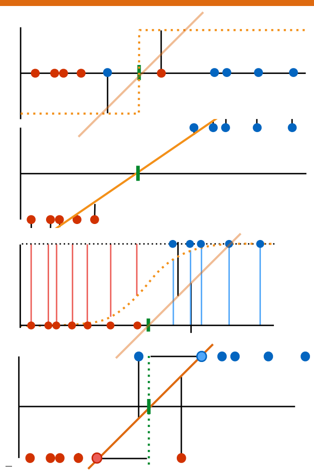
loss functions

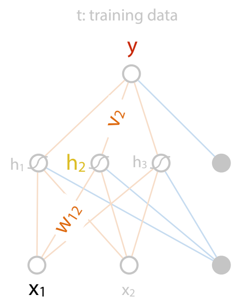
accuracy
aka zero-one loss, nr. of misclassified instances
doesn't work with gradient descent

least squares
assign -1, 1 to points treat as regression
doesn't really work well

cross-entropy
use a sigmoid to turn the linear output into probabilities
works well for non-non-separable data, overfits, has multiple solutions

soft margin SVM
aka Hinge loss, maximum margin loss
works well in high-dim data, separable data





$$l = (y - t)^2$$
$$y = v_1 h_1 + v_2 h_2 + v_3 h_3 + b_h$$
$$h_2 = \frac{1}{1 + \exp(-h_2')}$$
$$h_2' = w_{21} x_1 + w_{22} x_2 + b_x$$

$$\frac{\partial l}{\partial w_{12}} = \frac{\partial l}{\partial y}\frac{\partial y}{\partial h_2}\frac{\partial h_2}{\partial h_2'}\frac{\partial h_2'}{\partial w_{12}}$$

$$= 2(y - t) \cdot v_2 \cdot h_2(1 - h_2) \cdot x_1$$

$$w_{12} \leftarrow w_{12} - \eta \cdot 2(y - t) \cdot v_2 \cdot h_2(1 - h_2) \cdot x_1$$

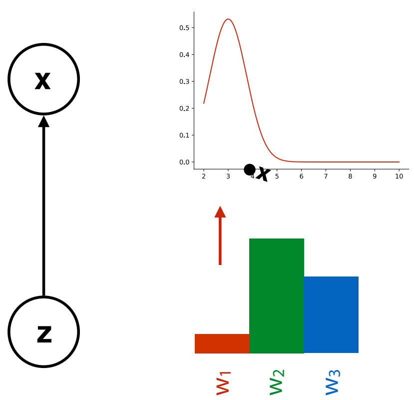
deep learning 1

Can we express the local derivatives in terms of the *tensor* inputs and outputs?

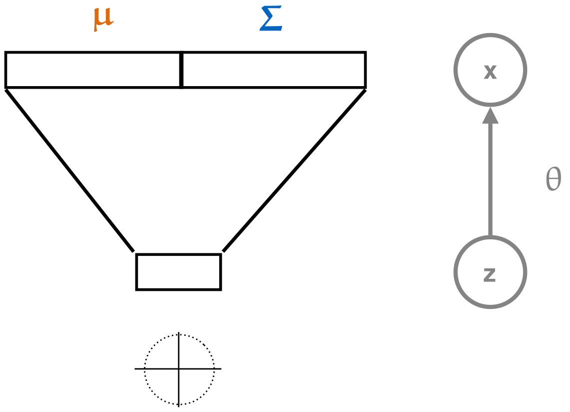
$$l = \sum (\mathbf{y} - \mathbf{t})^2$$
$$\mathbf{y} = \mathbf{V}\mathbf{h} + \mathbf{c}$$
$$\mathbf{h} = \sigma(\mathbf{k})$$
$$\mathbf{k} = \mathbf{W}\mathbf{x} + \mathbf{b}$$

$$\frac{\partial l}{\partial \mathbf{W}} = \frac{\partial l}{\partial \mathbf{y}}\frac{\partial \mathbf{y}}{\partial \mathbf{h}}\frac{\partial \mathbf{h}}{\partial \mathbf{k}}\frac{\partial \mathbf{k}}{\partial \mathbf{W}}$$

hidden variable model



deep learning 2



13

tree models

start with a single root node

loop until no unlabeled leaves

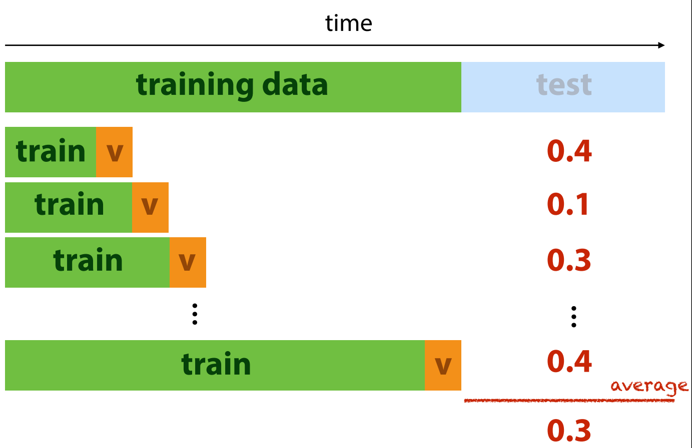
for each unlabeled leaf n with segment S:

if stop condition, label majority class of S

else split n on feature V with highest $I_S(V)$

14

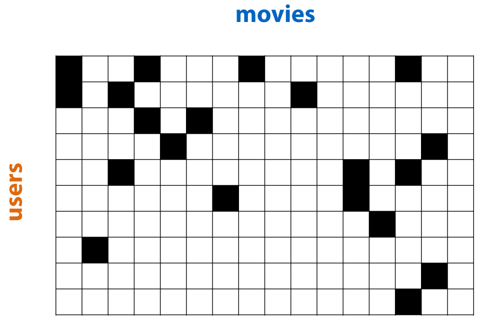
models for sequential data



15

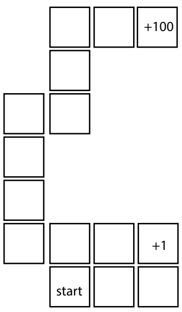
matrix models

ask users for ratings

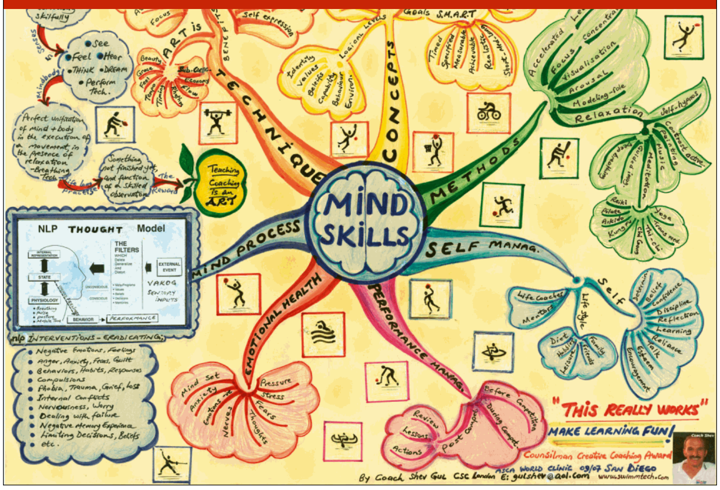


reinforcement learning

exploration vs. exploitation



mind maps



search

random search → gradient descent

simulated annealing

population methods

stochastic gradient descent

minibatch gradient descent

loss function
SVM
least squares
cross-entropy

EM

models

neural network

linear regression
linear classification
logistic regression

generator network

GAN

VAE — GMM — normal distribution

basic setting	sequences	matrices	online
features instances target values	separate sequences as instances or data as one big sequence	recommender systems graph models	reinforcement learning: agent-in-environment

deep learning:

no feature extraction
complex, deep models and backprop

probability

discriminative vs. generative models

frequentist vs. Bayesian approaches

hidden variable models
Expectation Maximization

Bayes' rule
prior
posterior
likelihood

maximum likelihood principle

joint, marginal, conditional

(cross) entropy, KL divergence

VAEs vs. GANs

mode collapse

Bayes optimal classifier

Bayes classifier

Naive Bayes classifier

smoothing

(multivariate) Normal Distribution

tricks of the trade

test set, validation set, training set

regularizers
- L1
- L2
- Dropout

normalisation
- linear
- whitening
- PCA

imputation

outlier removal

hyperparameter selection

weight initialisation

boosting, bagging, gradient boosting

abstract tasks

classification
linear, decision trees, logistic regression, kNN

regression
linear, regression trees, NN regression

probability/density estimation
MVNs, mixture models, Markov models, Naive Bayes

generative modeling
GANs, VAEs, Markov models, LSTMs

dimensionality reduction
PCA, Matrix factorization, Autoencoders

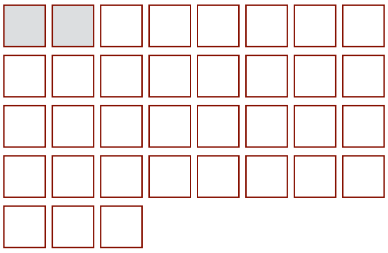
Recommendation/Collaborative filtering
Matrix factorization

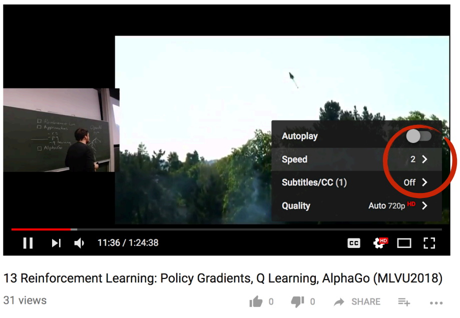
Reinforcement Learning
Q Learning, policy gradients

the exam

40 questions, multiple choice 4 answers each
in three categories:

- Recall (~ 33%)
- Combination (~33%)
- Application (~33%)
  all stuff we've practiced in the homework/practice exams





procrastination

NETFLIX

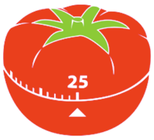
procrastination

Find the smallest viable commitment

Get an overview

Kill your perfectionism

pomodoro technique



let’s pretend the exam starts in…

10 minutes

30 minutes

tomorrow

exam strategy

Focus on the LECTURES, not the reading

Read the slides before watching the videos

Focus on the first 10 lectures

Make quick passes over everything. Figure out *what* you don't understand, then move on.

studying tricks

Compose a keyword list

Pages > Terminology

Come up with your own exam questions

Make random combinations

focus on the ins and outs

$$
\begin{aligned}
\nabla \mathbb{E}_a r(a) &= \nabla \sum_a p(a) r(a) \\
&= \sum_a \nabla p(a) r(a) \\
&= \sum_a p(a) \frac{\nabla p(a)}{p(a)} r(a) \\
&= \sum_a p(a) \nabla \ln p(a) r(a) \\
&= \mathbb{E}_a r(a) \nabla \ln p(a)
\end{aligned}
$$

$$\nabla \ln(z) = \frac{1}{z} \nabla z$$

Home
Announcements
Syllabus
Pages
Assignments
People
Discussions
Grades
Conferences
Collaborations
Modules
Quizzes
Files
Outcomes
Settings

Recommended reading

These materials are not required to pass the exam. But they are worth looking into
you just want to learn more.

Introduction

- A few useful things to know about machine learning. Pedro Domingos.
- Machine Learning crash course. From Google.
  - The glossary may be particularly useful if you get stuck on an unfamiliar

Linear Models 1

- A good way to visualize squared error loss.

Methodology 1

- Derived features: https://developers.google.com/machine-learning/crash-cours

---

Machine Learning Crash Course

OVERVIEW COURSE EXERCISES GLOSSARY

Introduction
Prerequisites and Prework

ML Concepts
Introduction to ML (3 min)
Framing (15 min)
Descending into ML (20 min)
Reducing Loss (60 min)
First Steps with TF (60 min)
Generalization (15 min)
Training and Test Sets (25 min)
Validation (40 min)
Representation (65 min)
Feature Crosses (70 min)
Regularization: Simplicity (40 min)
Logistic Regression (20 min)
Classification (90 min)
Regularization: Sparsity (45 min)
Introduction to Neural Nets (55 min)
Training Neural Nets (40 min)
Multi-Class Neural Nets (50 min)
Embeddings (80 min)

Introduction to Machine Learning

This module introduces Machine Learning (ML).

Estimated Time: 3 minutes

Learning Objectives

- Recognize the practical benefits of mastering machine learning
- Understand the philosophy behind machine learning

Introduction to Machin
Learning

Machine Learning Crash Course

OVERVIEW COURSE EXERCISES GLOSSARY

Machine Learning Glossary

This glossary defines general machine learning terms as well as terms specific to Tens

A

A/B testing

A statistical way of comparing two (or more) techniques, typically an incumbent agains
testing aims to determine not only which technique performs better but also to underst
difference is statistically significant. A/B testing usually considers only two techniques
measurement, but it can be applied to any finite number of techniques and measures.

accuracy

The fraction of predictions that a **classification model** got right. In **multi-class classific**
is defined as follows:

Correct Predictions

ttps://developers.google.com/machine-learning/crash-course/glossary

Seeing Theory

EN

Chapter 1

Basic Probability

This chapter is an introduction to the basic concepts of probability theory.

Chance Events

Expectation

Variance

2018 - P 4

- Home
- Announcements
- Syllabus
- Pages
- Assignments
- People
- Discussions
- Grades
- Files
- Quizzes
- Outcomes
- Collaborations
- Conferences
- Modules
- Settings

All

Search by title or author...

Pinned discussions

Terminology and notation
All sections
Last post at 24 Feb at 15:59

Typos and other small mistakes
All sections
Last post at 7 Mar at 11:31

Discussions

Questions exam 2018
All sections
Last post at 20 Mar at 17:04

Hw6, Decision Trees question 3
All sections
Last post at 18 Mar at 16:26

Project baseline requirements

break

what can't we do yet?

Causality

Compositionality

Generalization

causality

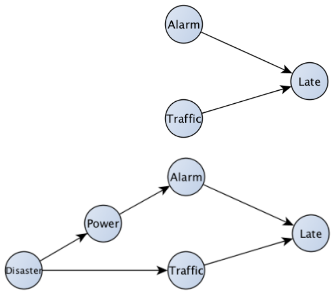
correlation does not imply causation

offline learning can only find correlations.

identifying causation requires intervention
i.e. a controlled experiment

causality without experiments

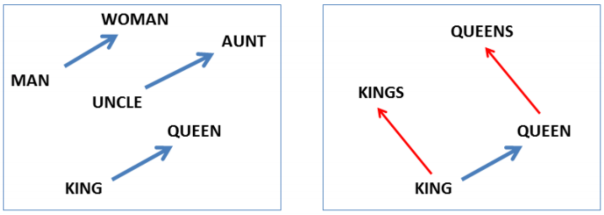
background knowledge



source:**https://medium.com/causal-data-science/if-correlation-doesnt-imply-causation-then-what-does-c74f20d26438**

compositionally

v(king) + $\underbrace{\text{v(woman) - v(man)}}_{\text{“feminine” vector}}$ ≈ v(queen)



(Mikolov et al., NAACL HLT, 2013)

generalisation

What if your test data is a little different from your training data.

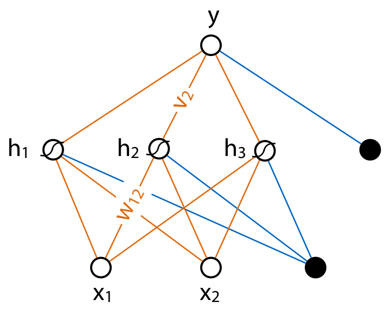
For instance:

train an RNN to sum numbers between 1 and 10
test on numbers between 1 and 15

causality, compositionality, generalisation

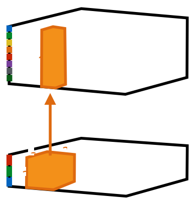
The key is to create a model with the right **inductive bias**

inductive biases: MLP



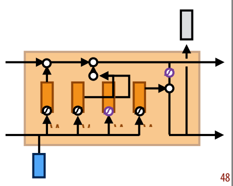
inductive biases of CNNs

- The data has a grid structure
  we know the data consists of pixels
- Inputs far apart on the grid are not relevant for low-level features
  we connect only a local group of pixels to each hidden node
- low level feature extractors are translation invariant
  we re-use the same weights for each patch



inductive biases of LSTMs

- The data is a sequence
- Each token can be modelled as a result of the tokens preceding it.
- Many tokens can be forgotten, and we can infer this from the token itself, together with the immediate context.



inductive bias

causality: inject background knowledge as an inductive bias

compositionality: add preference for compositionally explicitly, or model the rules of composition

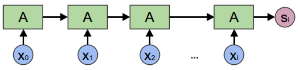
generalization: the more we constrain our model, the better it generalizes
but the less robust it is against the thing we didn't model

Grand challenge: start with the inductive bias, and let the model follow.

end-to-end learning

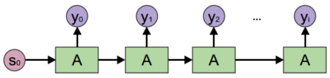
OCR → NER → REX → Rea

software 2.0 / differential programming



fold = Encoding RNN

Haskell: `foldl a s`



unfold = Generating RNN

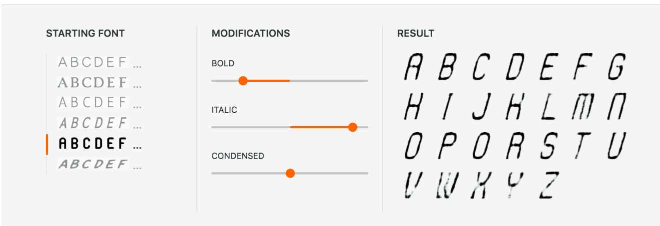
Haskell: `unfoldr a s`

medium.com/@karpathy/software-2-0

colah.github.io/posts/2015-09-NN-Types-FP/

Using generative models to invent meaningful creative operations

Let's look at an example where a machine learning model makes a new type of interface possible. To understand the interface, imagine you're a type designer, working on creating a new font [1]. After sketching some initial designs, you wish to experiment with bold, italic, and condensed variations. Let's examine a tool to generate and explore such variations, from any initial design. For reasons that will soon be explained the quality of results is quite crude; please bear with us.



distill.pub/2017/aia/

FACES

Smiling

Images from *Sampling Generative Networks* by White [6].

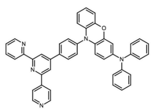
SENTENCES

"It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife."

Length

Sentence from *Pride and Prejudice* by Jane Austen. Interpolated by the authors. Inspired by experiments done by the novelist Robin Sloan [7]

MOLECULES



Delayed fluorescence decay rate

Images from *Automatic chemical design using a data-driven continuous representation of molecules* by Gómez-Bombarelli *et al* [8].

AI
artificial intelligence

IA
intelligence augmentation

II
Intelligent Infrastructure

GPT-2

SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Turkish | English | Spanish | Detect language → English | Spanish | Arabic | Translate

O bir doktor. O bir hemşire.

28/5000

He is a doctor. She is a nurse.

BUSINESS INSIDER

The first woman who appears in a Google Image search for 'CEO' is Barbie

Taylor Lorenz Apr. 10, 2015, 11:18 AM

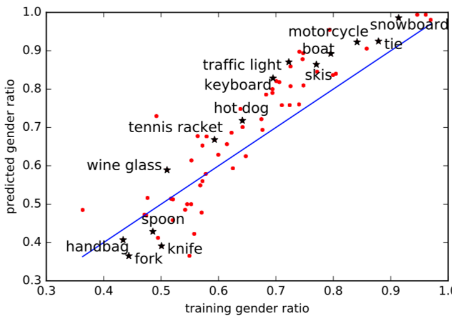
The first female result in a Google Image search for "CEO" is Barbie and is hidden at least 10 rows down, a recent post from The Verge points out.

CEO Barbie The Onion

And in fact, it's not even the real Barbie. As T.C. Sottek at the Verge notes, the image of Barbie in a power suit is from a from a 2005 Onion article stating that "women don't run companies," they just "work behind the scenes to bring a man's vision to light."

machine learning can *amplify* data bias

COOKING	
ROLE	**VALUE**
AGENT	WOMAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN



Film fans see red over Netflix 'targeted' posters for black viewers

The streaming service's customers say they are being duped by marketing that shows minor cast members as leading characters

▲ Set It Up is made to look like a two-hander between Taye Diggs and Lucy Liu, rather than the white couple. Photograph: Twitter Kelly Quantrill @codetrill

For anyone familiar with spending an unrelaxing hour scrolling through the Netflix menu trying to work out what to watch, the idea that one of the
/20/netflix-film-black-viewers-personalised-marketing-target#img-1 …onalise your viewing choices

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

What if

the data is a fair representation of the population

and

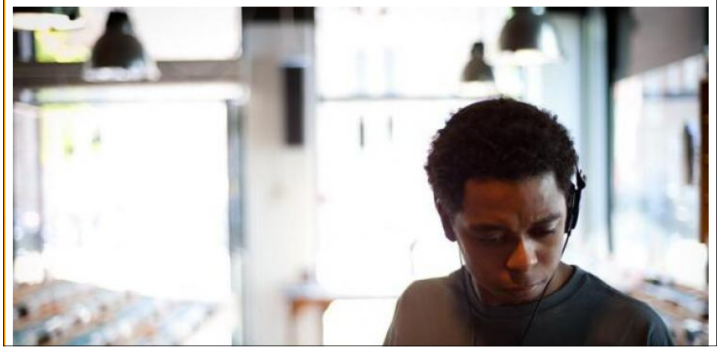
the predictions are accurate?

racial profiling

TOP STORIES

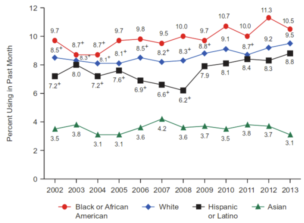
POLICE RACIAL PROFILING OVERWHELMINGLY APPROVED BY DUTCH PUBLIC

By Janene Pieters on June 6, 2016 - 09:02



drugs and race

Figure 2.12 Past Month Illicit Drug Use among Persons Aged 12 or Older, by Race/Ethnicity: 2002-2013



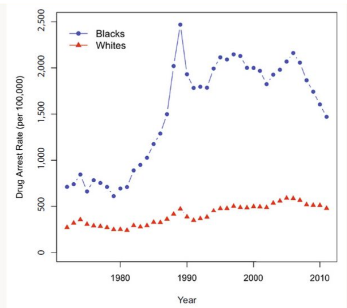


FIGURE 2-13 Drug arrest rates for blacks and whites per 100,000 population, 1972 to 2011.
SOURCES: Uniform Crime Reports race-specific arrest rates, 1980 to 2011 (accessed from BJS). 1972 to 1979 is taken from Federal Bureau of Investigation (1990).

prosecutor's fallacy

Abusing conditional probability

p(black | drugs) vs. p(drugs | black)

The probability that a basketball player is tall is different from the probability that a tall person plays basketball.

what if we forbid racial profiling?

Disallow the use of gender, ethnicity, sexual orientation etc. as features in sensitive ML tasks

What about: postcode, hobbies, average salary, mode of transport, etc.

What about companies: how do we police Google, Facebook, Yahoo?

What if

the data is a fair representation of the population

and

the predictions are accurate

and

we've correctly used Bayes rule?

actions versus predictions

It is fundamentally unfair to hold an individual responsible for the the actions of others that share their attributes.

Everybody has the right to to be judged on their own actions.

hold responsible:

subject to a traffic stop, not give parole, search at an airport, not give a credit card, make it more difficult to get a job.

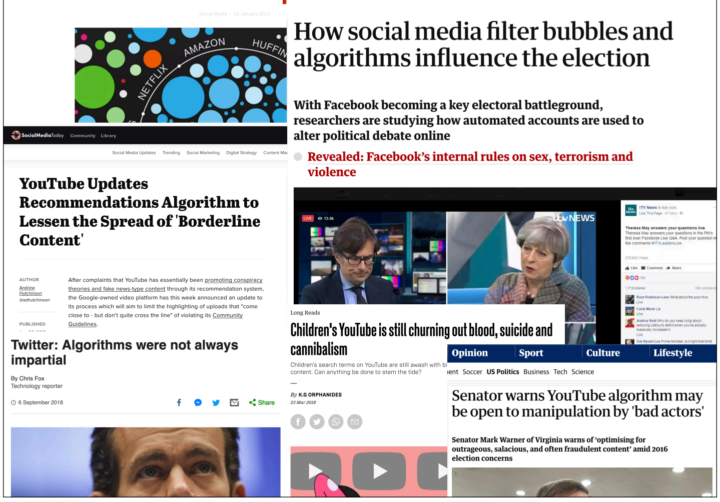
feedback loops

Offline learning doesn't stay offline.

Predictions become actions, that reinforce existing from the data.

It's not just about whether the predictions are accurate. It's about whether the actions are fair, and effective.

the problem of scale



WIRED

Technology | Science | Culture | Gear | Business | Politics

China

The complicated truth about China's social credit system

China's social credit system isn't a world first but when it's complete it will be unique. The system isn't just as simple as everyone being given a score though

By **NICOLE KOBIE**

21 Jan 2019

thank you for your attention

mlcourse@peterbloem.nl